Environmental Health Perspectives
Vol. 54, pp. 333–340, 1984

Ecologic Regression Analysis and the Study of the Influence of Air Quality on Mortality

by S. Selvin,* D. Merrill,† L. Wong,† and S. T. Sacks‡

This presentation focuses entirely on the use and evaluation of regression analysis applied to ecologic data as a method to study the effects of ambient air pollution on mortality rates. Using extensive national data on mortality, air quality and socio-economic status regression analyses are used to study the influence of air quality on mortality. The analytic methods and data are selected in such a way that direct comparisons can be made with other ecologic regression studies of mortality and air quality. Analyses are performed by use of two types of geographic areas, age-specific mortality of both males and females and three pollutants (total suspended particulates, sulfur dioxide and nitrogen dioxide). The overall results indicate no persuasive evidence exists of a link between air quality and general mortality levels. Additionally, a lack of consistency between the present results and previous published work is noted. Overall, it is concluded that linear regression analysis applied to nationally collected ecologic data cannot be used to usefully infer a causal relationship between air quality and mortality which is in direct contradiction to other major published studies.

Introduction

The relationship of air quality to disease has been extensively studied and still remains a fundamental health issue. One popular approach to studying this relationship combines measures of pollution and rates of mortality by means of a linear regression analysis applied to a series of defined geographic areas. Over a dozen research efforts fall into this class (*1,2*) [see Ricci (*3*) for a complete review]. The present work attempts to use regression techniques to reproduce the results of others, particularly the work of Lave and Seskin (1) and Mendelsohn and Orcutt (*2*) employing more extensive data than was previously available.

The assumptions and techniques of ordinary regression analysis are described by various authors (*4*). The classic regression analysis postulates that a dependent variable is linearly related to a series of independent variables, all of which are measured on the same observational unit. However, data are often not available on a unit basis but exist as statistical summaries of collections of units such as means, medians, percentages and rates. A natural extension of classic regression techniques is to analyze these collections of units without regard to the aggregated nature of the data, sometimes called ecologic regression analysis. The typical ecologic approach to the study of the influence of pollution on disease consists of analyzing a series of geographic units, using total mortality rates, air quality data and census-derived socioeconomic variables. Whether these types of data can be usefully employed to study air quality and health, and whether an ecologic regression model adequately reflects the complex relationships under study, are open questions.

Data and the Linear Regression Model

Three government agencies, which are required to routinely collect specific types of information, provided the principal data used to assess the influence of air pollution on health. Mortality data were tabulated from death certificate files at the National Center for Health Statistics (NCHS). Air quality data were extracted from records maintained by the Environmental Protection Agency in the SAROAD (Storage and Re-

*University of California, School of Public Health and Lawrence Berkeley Laboratory, Berkeley, CA 94720.
†Lawrence Berkeley Laboratory, Berkeley, CA 94720.
‡University of California Medical Center and Lawrence Berkeley Laboratory, Berkeley, CA 94720.

trieval of Aerometric Data) system. Each county in the United States was characterized with the use of variables from the U.S. County and City Data Book (*5*) of the U.S. Census Bureau. Some data concerning elevation and weather patterns were obtained from another source (*6*).

The smallest possible common geographic area for comparing and combining the three sources of data is the county, since the NCHS mortality data contain only the county of residence for each death certificate. This aggregation produced 3082 county records. (Not all government agencies use exactly the same county definitions. For example, independent cities in Virginia are sometimes considered separately and sometimes included with adjacent counties.) These counties form the ecologic unit for the analyses presented here. In addition, it was desirable to analyze these same data at a geographic level other than the county, primarily to compare our results with those of previous investigators (*2*). The second geographic level chosen was the 1970 Census Public Use Sample (PUS) area, in aggregation of counties.

The geographic units of the PUS are groups of counties which are fairly homogeneous with respect to socioeconomic status, and which divide the continental U.S. into 410 areas. Each area is so defined that its population exceeds 250,000 residents. For example, many large and sparsely populated counties in the western states like Montana, Colorado, Nevada, and Utah are aggregated to form single PUS areas, whereas large urban counties such as Los Angeles, Cook (Chicago), St. Louis and Baltimore are themselves PUS areas. A complete description of the Census Public Use Sample is available (*7*).

An average annual mortality rate for each county was calculated by taking the number of deaths in each county for the period 1968 through 1972 (only half the deaths were recorded in 1972) and dividing by 4.5 times the 1970 county population. Rates, for sex–, race– and age–specific categories were similarly calculated. The analyses focused primarily on total mortality rates, rather than cause-specific rates, so that direct comparisons could be made with the other major ecologic investigations of air quality and mortality. Mortality rates are subject to bias from a variety of sources, and these biases have been adequately discussed elsewhere (*8,9*).

Seventeen variables reflecting the 1970 socioeconomic status of U.S. counties, four variables concerning county weather patterns, and one other variable (county elevation) serve as the measurements of variation associated with mortality rates that are not directly related to air pollution. A list of these "control" variables is found in Appendix A. The essence of a multivariate aproach is the isolation of effects of specific variables (i.e., air pollution) from the influences of variables not of primary interest ("control" variables). The 22 of the 25 variables listed in Appendix A serve this "control" function.

The air quality data consists of measurements of three pollutants—total suspended particulates (TSP), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). These values were extracted from data collected at 6625 monitoring stations operating during the three-year period 1974 through 1976. County-level air pollution estimates were interpolated from average values at individual monitoring stations. The air pollution estimates for the 410 PUS areas are population-weighted averages of the county level value. A detailed discussion of the air quality data and interpolation methods is provided elsewhere (*10*).

The analysis of the total mortality rates employing 22 "control" variables and three air pollution measurements is restricted to white males and white females 45 to 54 years of age. The analysis of other racial groups is not practical, since too few deaths occurred during the period 1968-72 to calculate stable county-level mortality rates nationwide. Analysis of other age-specific categories provides little additional information, since the 25 independent variables have the same values for each county or PUS level analysis regardless of the age category being considered. The only new information contained in an age-specific analysis comes from the age-specific mortality rates themselves, which obviously differ within a geographic area. However, the average annual age-specific mortality rates for the four age categories 35–44, 45–54, 55–64, and 65+ increase fairly linearly (actually geometrically), which implies that the age-specific analyses will differ very little in statistical significance.

The underlying structure of a regression analysis (*4*) applied to mortality data (the dependent variable) can be investigated by checking for violations of the assumptions. Statistical summaries of the male and female mortality rates for both county and PUS data sets are given in Table 1. The measure of skewness and kurtosis reflect the structure of the population under investigation and for normally distributed data have expected values of zero. The 99.5% critical values are also given in Table 1. As can be seen, the observed values of skewness and kurtosis in all four columns are large in the sense that they are extremely unlikely to represent random deviations from zero. The measures of skewness and

Table 1. Summary statistics for total mortality rates/100,000 males and females for U.S. whites 1969–1974 by county and PUS areas.[a]

	PUS		County	
	Males	Females	Males	Females
Mortality rate (45–54) per 100,000	895.8	447.4	935.2	435.0
Standard deviation	146.1	56.0	252.0	123.0
Skewness	0.73	0.71	1.21	0.37
(99.5% critical value)	0.23	0.23	0.01	0.01
Kurtosis	0.80	2.29	5.82	2.95
(99.5% critical value)	0.50	0.50	0.23	0.23

[a]From Public Use Sample (7).

kurtosis indicate that mortality rates for ages 45–54 are skewed to the right (skewness > 0), and the probabilities associated with extreme rates are larger than expected from normally distributed data (kurtosis > 0). However, the assumption of a normally distributed dependent variable is generally not very critical to a regression analysis. Inferences made from approximately normally distributed data such as these mortality rates are not likely to be extremely misleading. It should be noted that the assumptions about the structure of the dependent variable do not affect the estimates made from the regression analysis but rather influence the statistical interpretation of these estimates (e.g., significance probabilities or p values).

Data which satisfy the basic assumptions of a linear regression analysis produce residual values [i.e., the observed dependent variables minus the values predicted from the estimated regression equation $(y - \hat{y})$] that vary randomly about a mean of zero and have negligible relationships to the independent variables. Figures 1 and 2 illustrate in standard deviation units the residual values from the regression analyses of males plotted against the rank of county and PUS population sizes. The county-level analyses for both males and females yield residual values that appeared to be random deviations from a linear model except for the least populous rural counties, where extreme (beyond ± 2 standard deviations) residual values are observed. That is, no evidence exists to reject the hypothesis that a linear model adequately describes the county level mortality patterns, except for a few counties with the smallest populations. The PUS residual values showed no trends with population size, which also indicates that no strong evidence exists for violation of the basic linear regression assumptions for the 410 PUS areas. Extreme values (beyond ± standard deviations) occurred with

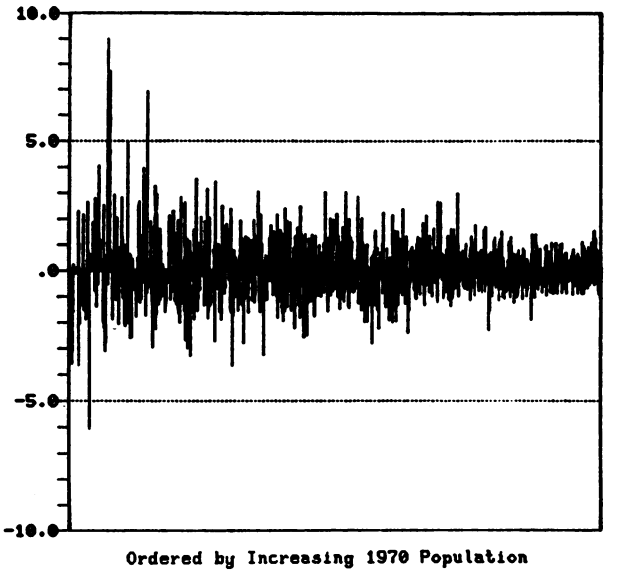

FIGURE 1. Residual for males, age 45–54, counties.

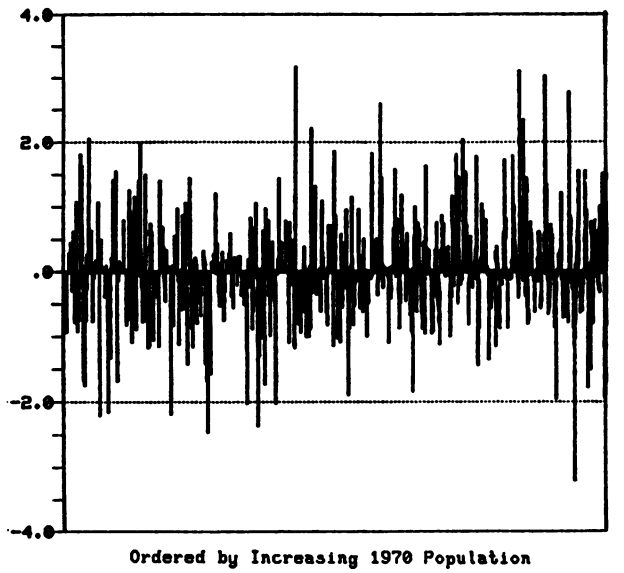

FIGURE 2. Residual for males, age 45–54, PUS areas.

expected frequency (about 5%) and had no geographic pattern for county and PUS regression analyses. Similarly, residual values plotted against other independent variables (not presented in detail here) also appeared to be randomly distributed. Since no strong evidence exists that mortality rates do not adequately fulfill the requirements for a dependent variable in a multiple regression analysis, transformations were not considered necessary, and each mortality rate was weighed equally in the following analyses.

Air Quality Regression Coefficients

If an ecologic regression analysis of air quality and mortality rates is meaningful, then the central issue is the interpretation of the regression coefficients associated with the air pollution measurements—TSP, SO_2, and NO_2. The combined influences of these three variables on the overall variation in nationwide total mortality rates is small but statistically significant both in the county and PUS analyses ($p = 0.009$ for males and $p < 0.001$ for females at the county level; $p = 0.035$ for males and $p = 0.010$ for females at the PUS level). The squared multiple correlation coefficient (R^2) increases from $R^2 = 0.312$ to $R^2 = 0.317$ (males) and from $R^2 = 0.105$ to $R^2 = 0.114$ (females) when the three air quality variables are added to the regression equation based on the 22 "control" variables for county data. Similarly, increases for the PUS analysis are $R^2 = 0.705$ to $R^2 = 0.712$ (males) and from $R^2 = 0.406$ to $R^2 = 0.432$ (females). Although air quality has a statistically significant influence, its magnitude is small. To gauge the "size" of this effect, it is helpful to similarly assess the influence of two other sets of variables. For example, the variables (% ≤ $3,000 and % ≥ $15,000) when added to the regression equation increase R^2 from 0.304 to 0.317 (males; PUS level data). When the five weather-related variables are added to the analysis, the R^2 values show an increase from 0.692 to 0.712 for males in county level data, and from 0.300 to 0.317 for males in PUS level data.

The direct comparison of the estimated regression coefficients (Table 2) presents no special problems since all air quality measurements are in the same units (micrograms per cubic meter). The comparison of the three regression coefficients relating U.S. (first row) air pollution to male mortality rates reveals a consistent but somewhat confused picture. The regression coefficients associated with the SO_2 measurement indicate strong and statistically significant associations for both county and PUS analysis. However, three of four coefficients for TSP and NO_2 are negative (one of these is statistically significant, with $p < 0.02$) which implies that lower mortality rates are accompanied by higher levels of TSP and NO_2.

The same analysis for the female mortality rates yields similar associations between the U.S. pollutant measurements and mortality, although the regression coefficients are reduced (in absolute value) in all six cases. Again, the SO_2 coefficients are positive, and three out of the four TSP and NO_2 coefficients are negative.

The same analyses were repeated employing a "dummy variable to account (in a limited sense) for regional variation. For example, TSP in the western counties typically contains a higher level of dust not found elsewhere. This type of regional variation is indeed expected and is incorporated in the regression equation by adding another independent variable [a somewhat simplistic solution to the issue but one which provides a direct comparsion with other work (2)]. This approach (rows 3 and 4 of Table 2) yields a reduction in the magnitude of the regression coefficients for most categories, but leaves the associations observed in the unadjusted national analyses essentially unchanged for all 12 coefficients. That is, SO_2 has a strong and positive association, and both TSP and NO_2 have predominantly negative associations, for county and PUS analyses for both sexes.

When the national data are stratified into four regional analyses (West = federal regions 8, 9 and 10; Midwest = federal regions 5 and 7; South = federal regions 4 and 6; Northwest = federal regions 1, 2, and 3) at a loss of some statistical power, no consistent pattern between air quality and mortality emerges. The SO_2 measurements show mostly positive associations with mortality rates with two exceptions ($b_i = -4.24$ for county data, West and $b_i = -14.36$ for PUS data, East). The coefficients associated with TSP and NO_2 are not consistent between county and PUS analyses

Table 2. Regression coefficients from 24 regression analyses of male and female mortality rates for TSP, SO_2 and NO_2 (county and PUS levels).

	TSP		SO_2		NO_2	
	Male	Female	Male	Female	Male	Female
U.S.						
County	−3.7*	0.4	11.3*	7.4*	−2.3	−2.0*
PUS	−2.6	−0.7	7.5*	6.5*	1.4	−0.3
U.S.[a]						
County	−2.4	0.8	7.1	4.1*	−2.6	−2.1*
PUS	−1.1	−0.5	4.3	3.9	1.2	0.0
West						
County	−7.2	9.2*	−4.2	2.5	10.1	3.6
PUS	2.5	−0.2	18.7	5.8	−5.6	−3.4
Midwest						
County	−1.4	2.6	1.1	− .3*	−6.8	6.5
PUS	2.5	2.4	−0.7	−1.6	3.1	1.7
South						
County	2.3	0.7	24.8*	2.0	−1.6	−1.1
PUS	−6.0	−0.6	28.2*	7.2	1.6	1.0
Northeast						
County	8.5	1.9	2.1	1.4	6.8*	−2.6
PUS	8.0*	−3.6	−14.4	5.0	2.1	2.8

*$p \leq 0.02$.

[a]Analyzed with a dummy independent variable to "account" for regional variation.

nor consistent between male and female analyses. Furthermore, as is the case with combined national data, many of the coefficients are negative and several of these are statistically significant.

Discussion

Like ionizing radiation, high levels of air pollution are unquestionably toxic. The existence of effects at low doses is equivocal. Parallel to the debate surrounding ionizing radiation, it is argued that most air pollution levels are low and mechanisms exist which protect against any disease effects. On the other hand, it is possible that no threshold level exists and any elevation of air pollution increases the risk of disease. Since large numbers of individuals are exposed daily to air pollution, evaluation of relevant data, methodologies and inferences bearing on the existence or nonexistence of a dose-response relationship between low levels of air pollution and disease risk is critical.

The results observed from both county and PUS regression analyses indicate that no persuasive evidence exists which links air pollution to mortality. Additionally, the analyses presented here do not repeat the results obtained by other authors (*1,2*). A multitude of minor reasons exist for the lack of agreement between the present regression analyses and those previously published. For example, different data sets were employed, different air quality interpolation methods were applied and different sets of "control" variables were used. However, regression techniques applied to similar data to investigate the same issue should produce reasonably consistent results before nationally collected data can be used to infer a causal relationship between air pollution and mortality. The reasons for this lack of consistency probably arise from three fundamental areas: the statistical issues underlying the use of linear models applied to aggregated data, the adequacy of ecologic data, and the interpretation of results (inferences) made from ecologic regression analyses. Analytic techniques, data and conclusions are indeed interdependent but detailed discussions of these issues are presented separately for clarity.

Statistical Issues

Regression analysis techniques are rigorously derived from a set of mathematical assumptions. The application of regression methods is less precise. The assumption of a normally distributed variable with equal variance linearly related to a series of independent variables is never completely realized. Some violations of these basic assumptions were detected in both the county and PUS data (e.g., mortality rates do not appear to be normally distributed). The analysis of residual values shows moderate deviations from the expected values but no large, nor clear-cut trends. The fact that relatively large quantities of data are available for each analysis (>1700 counties and >380 PUS areas) makes the analyses rather robust with respect to violations of some of the statistical assumptions. The questions of normally distributed data, equal variance, multicollinearity of coefficients, and adequacy of linear models could be investigated further but the present analyses indicate these purely statistical issues are not likely to be fundamentally important.

However, interpretation of the estimated regression coefficients is basic to the regression approach. The regression coefficients estimate the expected response in a dependent variable for a one unit change in an independent variable, while the other variables in the regression equation are held constant. The independent and dependent variables in an ecologic regression are summary values of aggregates of individuals. In this case problems arise in the interpretation of the regression coefficients when interest is focused on the individuals who make up the analyzed aggregate (*11*). That is, no straightforward nor bias-free interpretation exists of the ecologically derived regression coefficients with respect to the individual. The interpretation of ecologic regression analyses, in particular regression coefficients or correlation coefficients, as if they were derived from the classic regression assumptions, is often referred to as the "ecologic fallacy." Furthermore, both county and PUS regression analyses for males and females are attempts to estimate the same relationships; however many coefficients differed rather strikingly. Without a clear and consistent interpretation of the response of the dependent variable due to the isolated influences of specific independent variables (e.g., TSP, SO_2 or NO_2), the primary task of assessing the specific contributions to the variation in mortality rates from specific variables fails.

A statistical measure is declared significant if its value is unlikely to have occurred by chance variation. Analyses often yield statistically significant differences that have no consequential biologic influences, particularly when large amounts of data are involved. The county and PUS analyses of the influence of air pollution may fall into this category. Adding the three pollu-

tants (TSP, SO_2 and NO_2) to the regression equations produced a statistically significant increase in R^2 values. The evaluation of these increases from a biologic perspective is more difficult. It is entirely possible that increases like those observed (e.g., 0.005 for county and 0.007 for PUS–males) may be unimportant when assessed by other criteria. The question of "statistical" versus "biologic" significance is not unique to the study of air quality and disease, but should be kept in mind when evaluating the present results or those of other ecologic regression analyses.

Data Issues

Mortality data, particularly total mortality rates, are not ideal for epidemiologic analyses since they are subject to a variety of biases (*9*) and problems (*4,8*). The use of total mortality rates as a measure of risk minimizes the problems of statistical instability due to small numbers of observations and avoids biases due to disease classification. This choice also maximizes the squared multiple correlation coefficient calculated in a regression analysis. That is, general mortality patterns are more predictable from ecologic variables using a linear model than are age- or cause-specific rates. This increase in predictability is paid for by a decrease in biologic specificity. If an association is established for total mortality, the question immediately arises as to which of the many widely varying causes of death are in fact involved. The possibility also exists that employing an overall measure of mortality obscures important interactions among the specific causes of death. The age- and sex-specific rates produce a more epidemiologically focused analysis but are not accurately summarized by a linear model (R^2 low) (*10*). Total mortality leads to a high degree of predictability (R^2 high) but may yield relatively useless results since it is rare that a summary of a series of heterogeneous units is meaningful.

The 22 "control" variables present no technical problems. Furthermore sampling errors and biases in the PUS data are nonexistent or exist at extremely low levels. However, it should be emphasized that the important "control" variables are missing. Measures of cigarette smoking and occupational exposures are not included and are not tractable in the usual ecologic approach. The need to measure smoking in studies of air pollution has been pointed out by many investigators, most recently Holland et al. (*12*). The lack of smoking and occupational data in the ecologic approach is perhaps a fatal flaw.

The air quality data present concerns in several directions—coverage, exposure, and timing. Only 57% of the U.S. counties have adequate estimates of TSP, SO_2 and NO_2 based on a 60 kilometer criterion in the interpolation algorithm discussed elsewhere (*10*). These estimates vary in accuracy (monitoring density) but measure to some degree the air quality surrounding most of the nation's population. Although the coverage may be adequate, the degree of exposure of county residents is not directly measured for at least two reasons. Air quality monitoring stations are often placed to record specific sources of pollution and the data may or may not be representative of the area. For example, a station might be placed near a coal-burning utility company, so that air pollution measurements from this station would not generally reflect the actual levels experienced by the county residents. Secondly, neither mortality statistics nor "control" variables incorporate into the analyses the important aspects of population stability. The fact that a death certificate reports a person as a resident of a specific county does not necessarily imply that personal exposure levels are reflected by air pollution estimates for that county. An undetermined number of pesons will be new residents, or work elsewhere, or for a host of reasons, spend little time in the county of residence that appears on the death certificate. To the degree that this number is large, the county estimates will not accurately reflect exposure.

Whether the air quality measurements recorded in the EPA-SAROAD data base represent human exposure is one question. When the air quality was measured is another important issue. The present data involve mortality during 1968–1972 and air quality recorded during 1974–76. Other studies (*1,2*) are also forced to use rather recent air quality measurements since accurate nationwide data are available only for the last decade. For example, the work of Mendelsohn and Orcutt (*2*) used 1970 mortality and 1970 PUS data along with the 1974 air quality measurements. Implicit in analyzing mortality data from a time prior to the measurement of the air quality is the assumption that relative air quality differences among geographic units are stable over time. This important and usually ignored assumption implies that overall pollution levels could change, but relative differences must remain stable to be useful in an analysis of antecedent mortality rates. In fact, it is not obvious when the air quality measurements should ideally be made. If pollution affects mortality largely by increasing cancer rates, then air quality measurements should be made 10–20 yr prior to the mortality data, since this time interval is thought

to be the latency period for most cancers. Other causes of death have different latency periods and present a complicated picture for determining the ideal time to measure air quality.

Another potentially severe problem with geographically based variables used in ecologic regression analysis is that these variables are often averages of rather large and diverse units—counties, PUS areas or Standard Metropolitan Statistical Areas (SMSA) (*1*). Whether the ecologic variables are mortality rates, census summaries, or interpolated air quality measurements they are averages of potentially rather diverse observations. Analysis of this type of data does not address the basic concern that these averages may not accurately represent any specific quantity. That is, they are calculated from such a diverse set of measurements that they are relatively meaningless for understanding the nature of the relationships under investigation. For example, Los Angeles county has a population of over 7 million. Summaries such as median family income, percent black population, and percent owner-occupied homes may have little meaning, since these values ignore the many extremely different subpopulations (some of the nation's richest and poorest populations live in Los Angeles county). Los Angeles county is an extreme case, but to a lesser extent the averaging of heterogeneous observations into a single set of numbers occurs in all county-, PUS-, or SMSA-based data.

Inferences

The ecologic regression analyses of both county level and PUS level data sets produce neither strong nor consistent evidence that a link exists between ambient air pollution and mortality. A few associations (positive regression coefficients) do occur. The association between mortality rates and SO_2 levels is strong and positive for many analyses. The interpretation of this result is complicated. Taken at face value, a positive coefficient reflects a direct influence in terms of a linear model but the relationship between mortality and SO_2 is undoubtedly more complex.

Consider, for example, the observation that the coefficient associated with divorce rate is positive and in many cases significantly associated with total mortality. It is certainly simplistic to conclude that the divorce rate directly influences mortality. Although a regression equation is easily used to estimate the change in the number of deaths that would result from a specific percentage decrease in divorce (elasticity), this number would not be very plausible, nor would any corresponding estimates made from these ecologic regression equations. That is, decreasing the frequency of divorce alone is not likely to reduce mortality. Similarly, it is indeed possible that the positive association between SO_2 and mortality does not result from a direct causal relationship but rather from a complicated social/biological mechanism. Considering that TSP and NO_2 levels have mostly negative coefficients for a majority of analyses, the most likely explanation of any observed relationship between SO_2 and mortality is that the coefficients are artificially produced by the ecologic approach.

Protective effects (negative coefficients) from TSP and NO_2 pollutants are biologically implausible and result either from indirect associations with unmeasured variables (incomplete model bias) or are also strictly the result of the fallacy of drawing inferences from ecologically derived regression coefficients.

Appendix A

Sources of Air Quality and "Control" Data

The air quality and "control" data used in this analysis came from SEEDIS, the Socio-Economic Environmental Demographic Information System maintained at Lawrence Berkeley Laboratory. The data in SEEDIS came originally from three separate sources: items 1–17, 1977 County and City Data Book (*5*); items 18–22, 1977 Area Resource File (*12*); items 23–25, 1974–1976 Air Quality.

Control variables

1. Land area in square miles, 1970
2. Population, 1970
3. Net migration percent change, 1960-1970
4. Population, percent urban, 1970
5. Population, percent black, 1970
6. Population, percent foreign stock, 1970
7. Divorce rate per 1000 population, 1970
8. Persons, percent under 5 years, 1970
9. Persons, percent 65 years and over, 1970
10. Families, percent with income less than $3000, 1970
11. Families, percent with income $15,000 and over, 1970
12. Persons 25 year or more, percent with 4 years college or more, 1970
13. Employed, percent in manufacturing, 1970
14. Employed, percent in professional and mana-

gerial occupations, 1970
15. Occupied units, percent lacking some or all plumbing, 1970
16. Occupied units, percent with 1.01 or more persons room, 1970
17. Occupied units, percent owner occupied, 1970
18. January temperature, 1976
19. July temperature, 1976
20. January precipitation, 1976
21. July precipitation, 1976
22. Elevation, 1976

Air quality variables

23. Total suspended particulate, geometric mean concentration, of county at population centroid, micrograms per cubic meter, 1974–1976
24. Sulfur dioxide, geometric mean concentration, of county at population centroid, micrograms per cubic meter, 1974–1976
25. Nitrogen, dioxide, geometric mean concentration, of county at population centroid, micrograms per cubic meter, 1974–1976

The work described in this report was funded by the Electric Power Research Institute under Contract No. EPRI/DOE 790702/800410, and by the Office of Health and Environmental Research, Assistant Secretary for Environment of the U.S. Department of Energy under Contract No. W-7405-ENG-49.

REFERENCES

1. Lave, L., and Seskin, E. Air Pollution and Human Health. The Johns Hopkins University Press, Baltimore, 1977.
2. Mendelsohn, R., and Orcutt, G. An empirical analysis of air pollution dose-response curves. J. Environ. Econ. Manag. 6: 85–106 (1979).
3. Ricci, P. F., and Wyzga, R. E. A statistical review of cross-sectional studies of ambient air pollution and mortality. Paper presented at DOE Statistical Symposium, Berkeley, CA 1980.
4. Draper, H. R., and Smith, H. Applied Regression Analysis, Wiley, New York, 1966.
5. U.S. Bureau of the Census. County and City Data Book, 1977 (A Statistical Abstract Supplement). U.S. Government Printing Office, Washington, DC, 1978.
6. DHEW. The Area Resource File. U.S. Department of Health, Education and Welfare, Public Health Service, Health Resources Administration, Bureau of Health Manpower, Manpower Analysis Branch, DHEW Publication, No. (HRA) 80-4, October 1979.
7. U.S. Bureau of the Census. Public Use Sample Users Guide. U.S. GPO, Washington, DC, 1975.
8. Lilienfeld, A. M. Foundations of Epidemiology. Oxford Univ. Press, 1976.
9. Selvin, S., Sacks, S. T., and Merrill, D. W. Patterns of United States mortality for 10 selected causes of death. Lawrence Berkeley Laboratory Report LBL-10627, 1981.
10. Selvin, S., Sacks, S. T., Kwok, L., and Merrill, D. W. Ecologic regression analysis and the study of the influence of air quality on mortality. Lawrence Berkeley Laboratory Report LBL-12217.
11. Langbein, L. I., and Lichtman, A. J. Ecological Inference. Sage Publication, Beverly Hills, CA, 1978.
12. Holland, W. W., Bennett, A. E., Cameron, R., Florey, C. V., Leader, S. R., Schilling, R. S. F., Swan, A. V., and Walbar, R. E. Health effects of particular pollution: reappraising the evidence. Am. J. Epidemiol. 110: 527–659 (1979).
13. Winkelstein, W., and Kantor, S. Stomach cancer: positive association with suspended particulate air pollution, Arch. Environ. Health 18: 544–547 (1969).